QM Pins Food "Likes" and "Dislikes" With

Advanced Taste-Test Method

Small differences in similar foods, gross differences in checking general overall preferences, and group attitudes toward foods are now being quantitatively pegged using this hedonic scale adaption

DAVID R. PERYAM and NORMAN F. GIRARDOT
Quartermaster Food & Container Institute, Chicago

Reprinted from FOOD ENGINEERING
July 1952, pp. 58-61, 194

Approach to the Elusive

Attainment of reliability in consumer-preference evaluations of foods continues to be a troublesome problem, despite the various attempts at solution.

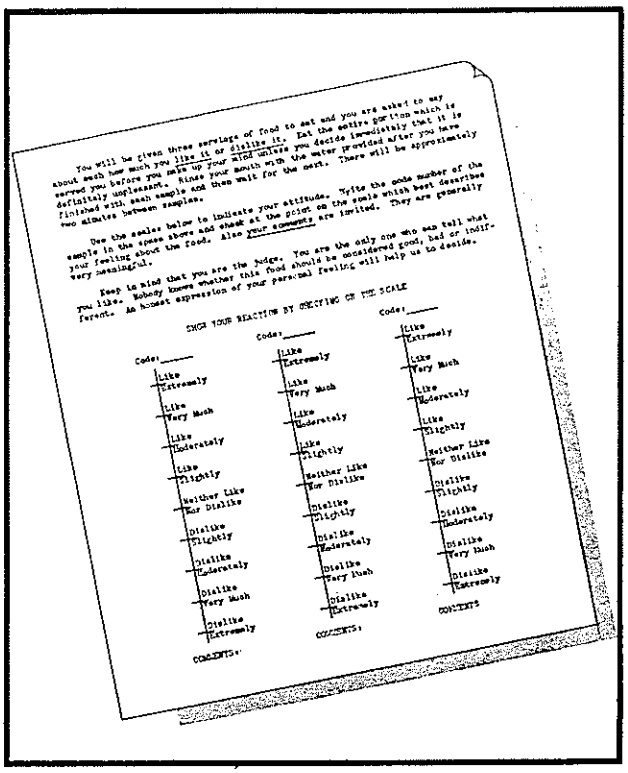
***First*, there has been general failure to achieve standardization.**

***Second*, this failure has fettered development of confidence on the part of potential testers.**

With a practical answer still elusive, the immediate logic is to entertain some partial solution that shows promise of lending stability.

Accordingly, the hedonic scale system is here advanced as a technique which may very well prove to supply this needed foundation for development of tomorrow's consumer preference methods.—*The Authors*

Lab Form Employed



QUESTIONNAIRE used in laboratory to judge specific food preferences with the hedonic scale method. Tester checks point which best describes his reaction to food.

QM Pins Food "Likes" and "Dislikes" With

Advanced Taste-Test Method

Small differences in similar foods, gross differences in checking general overall preferences, and group attitudes toward foods are now being quantitatively pegged using this hedonic scale adaption

DAVID R. PERYAM and NORMAN F. GIRARDOT
Quartermaster Food & Container Institute, Chicago

A technique has been developed at the Quartermaster Food & Container Institute which, we feel, offers notable progress in evaluating consumer preference of foods. It has been labeled the "hedonic scale method."

This basic approach is not new, since the method uses a variant of the well-known rating scale, introducing the hedonic value concept, which refers to the psychological range of "unpleasant" or "dislike" at the lower end to "pleasant" or "like" at the upper end.[1]

Also, the problem of quality control of food flavors is completely divorced from this technique—thus keeping the analysis of consumer preferences on a separate plane.

This type of scale was first tried at the Institute in a comparison of methods of predicting soldiers' food choices. Results were encouraging, and in 1949 its suitability for the study of relatively permanent preference attitudes toward food was demonstrated. It was also shown to be adequate for laboratory use in measuring the response to foods as eaten. Forms and procedures were developed for both situations, and since then both kinds of applications have been used constantly.

The hedonic scale method is not considered a polished system, because pertinent questions as to its interpretation, its reliability, and the extent of its usefulness are as yet unanswered. Unquestionably, it can be improved. But it is described here in its present form because of the interest that it has aroused among many people concerned with preference evaluation of foods.

To present the hedonic scale method in terms of a set of rigid specifications would misrepresent the

Field Form Names Foods, Can be Used Any Time

	FOOD ITEM	LIKE				INDIF-FERENT	DISLIKE			
Not Tried	Cream gravy	Like Extremely	Like Very Much	Like Moder-ately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moder-ately	Dislike Very Much	Dislike Extremely
Not Tried	Bread pudding	Like Extremely	Like Very Much	Like Moder-ately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moder-ately	Dislike Very Much	Dislike Extremely
Not Tried	cheese	Like Extremely	Like Very Much	Like Moder-ately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moder-ately	Dislike Very Much	Dislike Extremely
Not Tried	French fried onions	Like Extremely	Like Very Much	Like Moder-ately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moder-ately	Dislike Very Much	Dislike Extremely
Not Tried	Lettuce wedges	Like Extremely	Like Very Much	Like Moder-ately	Like Slightly	Neither Like Nor Dislike	Dislike Slightly	Dislike Moder-ately	Dislike Very Much	Dislike Extremely

IN FIELD, preferences are registered on forms like this one—usually with nine items to be reviewed at one time. This type check-sheet was designed for studying soldiers' preferences either a few minutes or several days after eating foods under question. In scoring, each scale point is given a number.

situation—since questionnaire forms can vary, and considerable latitude is allowable in the test. Forms and procedures now in use at the Institute will serve as a basis for this description, with the critical features being emphasized and the points of permissible variation indicated.

In the standard questionnaire form used for laboratory consumer preference evaluations, two main parts may be recognized—the instructions and the scales. Instructions are generalized, without reference to any particular food, and with provision for a maximum of three test items.

When the questionnaire is designed to measure general attitudes toward foods, the scale may be presented in a much different form for a larger number of items. However, all forms are the same in the following two respects: The phrases which describe the scale points do not change, and they are always placed so their continuity will be seen.

The method is designed for use with observers who are entirely without experience in food testing. Both the instructions and description of scale points are written with this purpose in mind. However, there is no evidence that the resulting simplicity reduces its effectiveness with other, more sophisticated observers.

The instructions have two functions: One is, to tell the observer what he must know, or what the experimenter wants him to know, about the mechanics of the test; the other, and more important function, is to encourage him to report his immediate naive response without any conscious effort to remember or to judge. The simple "like-dislike" description of the scale further encourages this tendency.

Oral instructions can be fully adequate, and they are desirable if the test situation permits individual contact with each observer as in the laboratory. But sometimes the only contact will be through the questionnaire, hence instructions must be carefully written.

No evidence so far available has shown that the geometry or arrangement factors of the scale are very critical, although these have not been intensively investigated. Adjustments are often made for such factors, how-

SOLDIERS test coffee in mess hall at Fort Bragg during a recent field test conducted by QM Board. Immediately after finishing, men rate each sample on hedonic scale questionnaire. (Department of Defense photo)

ever, in developing rating scales for other purposes.[2] Laboratory and field forms seem to work equally well. The vertical scale on the laboratory form is to give the idea of a continuum with equidistant points, but there is no proof that this has any definite effect. Apparently most observers consider the laboratory scale only as a series of categories—the same as in the field form.

The lab scale is normally presented with the "like extremely" category at the top or left. But reversal of both the horizontal and vertical scales has been tested and no definite differences have been found. Also, the physical size of the scales has varied from 5 to 7 in. Probably extreme variations in scale size would affect results, although this has not been experimentally determined.

The Taste Tester

Selection of test people is of great importance to interpretation of the results. However, the adequacy of the method does not depend too heavily upon this. The objective is to measure group responses toward foods. Whether or not the group tested represents the consumer group in which we are interested, or whether it represents any group at all, must be determined independently of the test itself.

The number of people required for a given test cannot be arbitrarily stated, but must be determined by the experimenter by the nature and importance of the problem and the degree of precision desired in the results. In the laboratory, the number potentially available usually influences that decision, too.

Variability in ratings from a group of observers tends to be high but also tends to be fairly constant. Thus, it becomes possible to estimate with some accuracy the number of observers necessary to assure statistical significance for a given scale-point separation between two test foods. Because of this high variability, the scale is not suited for use with very small panels—the standard number of observers for tests at the Institute being 40, although this may be increased for important problems. The number of respondents in a field test is usually determined by criteria other than the desired precision of the results.

The tester in the laboratory receives a maximum of three foods at any session. He is instructed to rate each food as he finishes it, and to rinse his mouth with water between samples. Further, he is asked not to change a rating once it has been made. This is done to encourage him to consider the foods independently of one another. Whenever more than one sample is presented it must be assumed that ratings may mutually influence each other.

It has been demonstrated that there is a definite contrast effect when foods being tested lie far apart on a scale.[3] Presenting only one sample at a session would prevent this but is wasteful of laboratory and observer time. Note that later samples cannot affect the earlier ratings except for the occasional observer who fails to obey instructions. Forward-acting effects are equalized by varying the order of presentation of the samples.

The laboratory test situation is designed for optimum sensory discrimination, and the observer's immediate impression is recorded with no opportunity for a memory lag. In field testing, this cannot be done too often. The respondent will usually eat the test food along with other foods as part of his normal meal, and then be asked to comment on it anywhere from a few minutes to several days later.

Thus, when general preference attitudes are surveyed, the delay factor becomes even more important, with the possibility that the respondent may answer from experiences which he could not remember even if he tried. Therefore results of laboratory tests have been found to be more reproducible than those of field tests.

Checking the Data

Two approaches are used in analysis of the data. Each results in a kind of "preference index."

In the first approach, numbers from 1 to 9 are assigned to the scale's nine categories and the data then treated quantitatively. The numbers may begin at either end of the scale, but to have high numbers reflect preference, 9 is usually assigned to the "like extremely" end. Score distributions are then dealt with by usual statistical procedures. Calculated are means, standard deviations, standard errors of the means, and the significance of differences between means. And both scores and means may be treated by analysis of variance.

The validity of using certain of the statistical methods with data of this type is questioned by some statisticians. But analytical methods are required and use of the normal procedures is justified on practical grounds until more appropriate techniques become available. The mean rating is the statistic most often used at the Institute, being reported as the major test result.

The second approach provides an index which is probably as useful as the first for the practical purpose of describing a group response to a food. Statistically, it is more respectable, since it deals only with the percentages of responses falling into the various categories. However, unless the number of observers is large, the percentage of responses in some of the categories may be zero, or very close to it. Then it is more convenient and meaningful to combine the categories.

One grouping which has self-evident validity is a combination of the four categories of "dislike." This statistic is always calculated by us and reported along with the mean rating. Also, the percentages falling into single categories, such as "like extremely" or "dislike extremely," may be useful in special analyses.

Results for four foods which were tested at different times in the laboratory are presented in Table I. The entire distribution of responses is shown, along with the statistics which are usually calculated. Food A's rating is unusually low, and that item would be considered nonacceptable under any circumstances; food D has one of the highest ratings ever obtained in the laboratory tests. Of the other two, B would be considered "poor" and C, "good."

The broad distributions of responses and the resulting high standard deviations that occur with all of the foods except D are typical. They do not necessarily indicate lack of precision in the method, but reflect the fact that there are normally wide differences among people in their feelings about foods.

Adequate Accuracy

Questions of both theoretical and practical importance should be asked about any new and relatively untried method. These concern its reliability, its precision of discrimination, and its validity for various purposes.

Experimental evidence in regard to the hedonic scale method is far from adequate, but inferences can be drawn from results obtained over its two-year period of use in our laboratory. This evidence suggests that the mean hedonic rating from as small a group as 40 observers will have satisfactory stability.

However, we may look at the question of stability in two different ways: First, how reproducible is an individual rating, or a mean rating, when a test is repeated under identical conditions?

Three ration items were rated by 35 observers and the test repeated three weeks later without their knowing that the same items were involved. The sets of individual ratings corre-

lated very well and all the mean ratings were reproduced within 0.2 scale points. Under these conditions there seems to be little question about the reproducibility of results.

Second is the point that for practical work, one wants to know what happens when conditions are not as carefully controlled. And here, even under ordinary conditions, mean ratings have been found to be stable enough to give the impression that a constant property of the food is being measured. Results will generally fall within the narrow range of values permitted by the experimental error.

Table II shows results obtained upon repeated testing of four foods by groups of 40 each. Some of the tests were widely spaced in time, the combinations of foods served were not constant, and the groups were never the same, although they were drawn from the same population. Presumably, the quality of the test foods remained the same, though slight variations may have occurred.

Statistical methods were used to find out whether the observed differences between ratings for a food were due to chance. The actual ratio of the range of mean ratings to the average standard error for each food is shown in the column, "Range/Standard, Actual" of the table. The expected value of this ratio for any given number of test repetitions, in situations where only chance is operating, may be calculated by statistical methods.[4] These values are shown in the column, "Range/Standard, Chance" of the table.

In only one case does the actual ratio appreciably exceed the chance ratio. This suggests that most differences will be accounted for as normal chance variation and that the test is inherently reproducible.

Precision of discrimination with a rating scale is determined by the scale distance between mean ratings, the variability within the distribution of individual ratings on each item, and the number of observers used in each test. If the scale is measuring hedonic value at all, and if there is a true difference in group response toward the test foods, the difference can be proven simply by increasing the number of observers as necessary.

What Scores Mean

Approximately, 2,000 tests run in the Institute laboratory on over 100 different items have shown a total range in mean ratings from 2.9 to 8.5. Generally, mean ratings below 5.0 represent either poor quality foods or foods that are strange to the observers, while those over 7.5 are obtained for good quality samples of highly popular foods, such as ice cream and candy.

Most foods fall in the range of 5.5 to 7.5, with variability among individual ratings tending to be high. The standard deviations shown in Table I for foods A, B, and C are typical. When variability is of this order and the observer group consists of 40 persons, differences in mean rating of about 0.8 scale units will usually be significant in the sense that they will be reproducible about 95 percent of the time.

Theoretically, then, at least six mutually exclusive levels of hedonic value could be established over the total range.

The hedonic scale rating reflects the attitudes of a group of people toward certain foods and under a given set of conditions. How well the observers and the test conditions represent any practical use situation will depend upon the adequacy of the test plan and the sampling procedures.

Since the hedonic scale method creates no unique problems in this regard and has no special limitations, the factors affecting its validity will not be discussed in detail. However, the ease with which the scale is understood by most people, and its fairly good observed reliability, suggest that in regard to validity for uses involving prediction of consumer preferences it will certainly be as good as other available methods.

Experience to date has shown certain purposes for which the hedonic scale method is valid if the sampling of observers is appropriate and the tests are properly run. These may be summarized as follows:

1. To detect small differences in the direct response to similar foods.

2. To detect gross differences in the direct response to foods, even when time, subjects, and test conditions are allowed to vary.

3. In field questionnaire surveys, to reveal differences in group-preference attitudes toward foods.

A fourth purpose may be included, but with some reservation. This is to make general predictions, on an absolute basis, about the acceptance level of any food.

But many people who are not familiar with the problems of preference measurement tend to consider the indices derived from this scale as if they were fixed and unchangeable indicators of acceptance. Available evidence does not bear this out. Even though ratings have a certain stability, they will vary with such factors as the psychological and physiological state of the consumer.

Obviously, they may also vary according to the type of consumer group tested. Thus, it is undesirable to try to establish or use fixed standards of interpretation. For example, mean ratings below 5.0 are usually obtained only for poor quality foods, but it cannot be categorically stated that this number marks the boundary between "acceptable" and "non-acceptable." Exceptions have been found where foods which rate below 5.0 show satisfactory field acceptance.

Another caution is also in order: The hedonic scale method cannot be considered for quality control of flavor in food production. Even though the method has been discussed only in relation to the measurement of preferences, this caution is believed necessary, since the two problems are not always recognized as being essentially different.

Two factors tend to disqualify the

TABLE I—DISTRIBUTIONS OF RESPONSES on Hedonic Scale, With Resulting Statistical Indices for Various Food Items

Scale Point Description	Assigned Value	Frequency of Responses			
		Food A	Food B	Food C	Food D
Like Extremely	9	0	0	9	11
Like Very Much	8	1	5	12	21
Like Moderately	7	1	11	7	8
Like Slightly	6	8	4	7	0
Neither Like Nor Dislike	5	3	3	2	0
Dislike Slightly	4	4	6	0	0
Dislike Moderately	3	6	6	0	0
Dislike Very Much	2	10	5	2	0
Dislike Extremely	1	7	0	1	0
Total Responses		40	40	40	40
Mean Rating		3.48	5.20	7.08	8.08
Standard Deviation		1.99	2.04	1.93	0.68
Percentage "Dislike" Responses		67.5	42.5	7.5	0.0

TABLE II—REPRODUCIBILITY of Hedonic Scale Results, With Groups of 40 Observers Repeatedly Testing From Foods at Different Times

Food	Number of Tests	Analysis of Mean Ratings				Range/Standard Error	
		Lowest	Highest	Range	Standard Error [1]	Actual [2]	Chance [3]
Fresh Milk	11	7.02	7.88	.86	.19	4.5	3.4
Lemonade	4	6.60	7.29	.69	.32	2.2	2.6
Canned Bread	8	6.09	7.05	.96	.25	3.8	3.2
Pea Soup	6	7.08	7.65	.57	.22	2.6	2.9

[1] Average of individual standard errors of the means.

[2] Figure of Column "Range" divided by figure of Column "Standard Error."

[3] Expected value of ratio if all differences between means were due to chance.

method for this type use: 1. Large test variations mean a considerable number of observers are required for precision. 2. The type of responses that are called for are expected to change with a number of conditions which cannot always be controlled. And to correct these would again require a larger number of observers in each test than generally are available for quality control work.

Research on this new method is continuing at the Institute with these two main objectives: Improvement of the method itself, and determination of the relative importance of the factors which affect hedonic responses toward food.

References

1. Beebe-Center, J. G., *Pleasantness and Unpleasantness,* D. Van Nostrand, New York, 1932.
2. Guilford, J. P., *Psychometric Methods,* McGraw-Hill, New York, 1936.
3. Hanson, H. L., Kline, L., and Lineweaver, H., "Application of Balanced Incomplete Block Design to Scoring of Ten Dried Egg Samples," *Food Tech.,* 5, No. 9. 1951.
4. Grant, E. L., *Statistical Quality Control,* McGraw-Hill, New York, 1940.

The above article is taken from the paper, "The Hedonic Scale Method of Measuring Food Preferences," designated as No. 385 in the series reporting research of the Quartermaster Food & Container Institute for the Armed Forces. The Institute points out that the views presented are those of the authors and are not to be construed as necessarily reflecting endorsement of the Department of Defense.